The Oedipe system at CLEF-QA 2005

Romaric Besançon, Mehdi Embarek and Olivier Ferret
CEA-LIST
LIC2M (Multilingual Multimedia Knowledge Engineering Laboratory)
B.P.6 - F92265 Fontenay-aux-Roses Cedex, France
{besanconr,embarekm,ferreto}@zoe.cea.fr

Abstract

This article presents Oedipe, the question answering system that was used by the LIC2M for its participation to the CLEF-QA 2005 evaluation. The LIC2M participates more precisely to the monolingual track dedicated to the French language. The main characteristic of Oedipe is its simplicity: it mainly relies on the association of a linguistic pre-processor that normalizes words and recognizes named entities and the principles of the Vector Space model.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing — Linguistic Processing; H.3.3 Information Search and Retrieval — Search process; H.3.4 Systems and Software — Performance evaluation (efficiency and effectiveness); H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, Monolingual question answering, Minimal question answering system

1 Introduction

Question Answering is at the edge of Information Retrieval and Information Extraction. This position has led to the development of both simple approaches, mainly based on Information Retrieval tools, and very sophisticated ones, such as [8] or [7] for instance, that heavily rely on Natural Language Processing tools. Previous evaluations in the Question Answering field have clearly shown that high results cannot be obtained with too simple systems. However, it still seems not clear, or at least it is not a shared knowledge, what is actually necessary to build a question answering system that is comparable, in terms of results, to the best known systems. This is why we have decided to adopt an incremental method for building Oedipe, the question-answering system of the LIC2M, starting with a simple system that will be progressively enriched. Oedipe was first developed in 2004 for the EQUER evaluation [1] about question answering systems in French. It was designed mainly for finding passage answers and its overall design was not changed for its participation to the French monolingual track of CLEF-QA 2005. The main adaptation we did for CLEF-QA was the addition of a module that extracts short answers in passage answers for definition questions.

2 Overview of the Oedipe system

The architecture of the Oedipe system, as illustrated by Figure 1, is a classical one for a question answering system. Each question is first submitted to a search engine that returns a set of documents. These documents first go through a linguistic pre-processor to normalize their words and identify their named entities. The same processing is applied to the question, followed by a specific analysis to determine the type of answer expected for this question. This search is performed through three levels of gisting: first, the passages that are the most strongly related to the content of the question are extracted from the documents returned by the search engine. Then, the sentences of these passages that are likely to contain an answer to the question are selected. These sentences can also be considered as passage answers. Finally, minimal-length answers are extracted from these sentences by locating their phrases that best correspond to the question features.

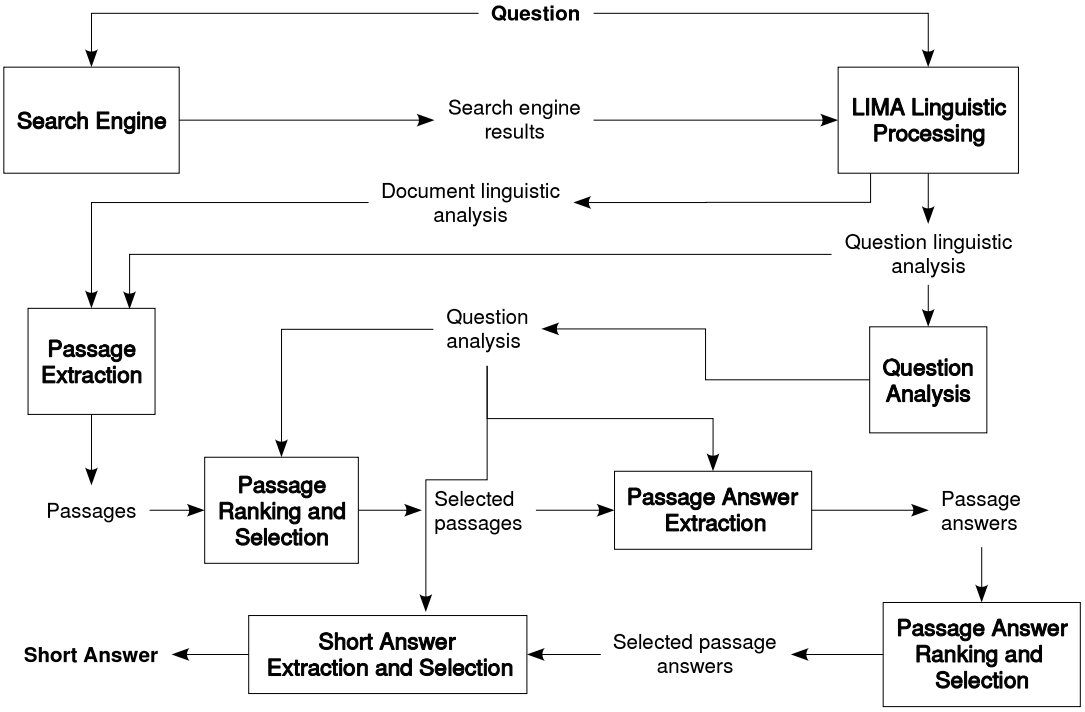

Figure 1: Architecture of the Oedipe system

3 From documents to passages

3.1 LIMA

LIMA [3], which stands for LIc2m Multilingual Analyzer, is a modular linguistic analyzer that performs text processing from tokenization to syntactic analysis for 6 languages[1]. More precisely, for CLEF-QA, the linguistic analysis of both documents and questions relied on the following modules:

- tokenizer

[1] These languages are: French, English, Spanish, German, Arabic and Chinese. Full syntactic analysis is only available for French and English but the chunker module exists for the other languages.

- morphological analysis
- detection of idiomatic expressions
- part-of-speech tagging
- content word identification
- named entity recognition

We decided not to use the module of LIMA for extracting compound nouns as previous experiments on TREC data had shown that for question answering, such compounds are useful for selecting documents [6] but are not necessarily interesting for selecting sentences where an answer can be found: there is no reason for answers for being systematically at a short distance of a compound of their question. Moreover, the positive effect of compounds for the selection of documents was already obtained through the search engine we used.

3.2 Search engine

For the first selection of documents from the collection, we used the LIC2M search engine, that had already participated to the Small Multilingual Track of CLEF in 2003 [4] and 2004 [5]. This search engine is concept-based, which means that it focuses on identifying in a query its most significant concepts, generally represented as multi-terms and named entities, and favors in its results the documents that contain one occurrence of each query concept, or a least the largest number of these concepts, whatever their form[2] and their number of occurrences in documents. The search engine relies on LIMA for the linguistic analysis of both the documents and the queries. The configuration of LIMA was the same as the one described in the previous section, except that the compound extractor was added.

For CLEF-QA, no specific adaptation of the search engine was done. Each question was submitted to the search engine without any pre-processing and the first 50 documents given as result were selected for the next steps.

3.3 Question analysis

The analysis of questions aims at determining the type of the expected answer. More specifically, it determines if the answer is a named entity or not, and in the first case, the type of the target named entity. We distinguish only 7 types of named entities: person, organization, location, date and time, numerical measure, event and product. Except for the two last ones, they correspond to the types of named entities defined by the MUC evaluations. The analysis of questions is achieved by a set of 248 rules implemented as finite-state automata. These automata are similar to those defined for recognizing named entities and idiomatic expressions in LIMA. Each rule is a kind of lexico-syntactic pattern that can also integrate semantic classes. When it is triggered, it associates to the question one type among the 149 question types we distinguish. As this typology heavily relies on the surface form of the questions, a mapping is defined between the question types and the answer types. A question can have several answer types when the rules are not sufficient for choosing among them. This is the case for some ambiguities between persons and organizations.

The rules for question analysis were elaborated following a semi-automatic method, first developed for the EQUER evaluation [2] and that is inspired by Alignment-Based Learning [9]. This method starts from a corpus of questions, made in our case of translated TREC-QA questions[3] and questions from the previous CLEF-QA evaluations. First, the edit distance of Levenshtein is computed for each pair of questions. Then, the Longest Common Substring algorithm is applied

[2] The search engine can recognize a concept of the query if it appears in a document as a synonym or a subterm.

[3] More precisely, these questions come from the TREC-8, TREC-9 and TREC-10 evaluations and were translated by the RALI laboratory.

for each pair of questions that are close enough in order to extract their common part. The common substrings are sorted to find the question types whereas the distinct parts can be grouped to form classes of entities with similar characteristics:

What is the capital of Yugoslavia?
What is the capital of Madagascar?

question type: what_is_the_capital_of
class of countries: Yugoslavia, Madagascar

This method is implemented by the CoPT tool, developed by Antonio Balvet[4].

3.4 Passage extraction, ranking and selection

After the selection of a restricted set of documents by the search engine, Oedipe delimits the passages of the documents that are likely to contain an answer to the considered question. This delimitation relies on the detection of the areas of documents with the highest density of words of the question. It is done by giving to each position of a document an activation value: when such a position contains a word of the question, a fixed value is added to its activation value and to the activation value of the positions around it (*activSpread* positions on the right and the left sides). Finally, the delimited passages correspond to the contiguous positions of the document for which the activation value is higher than a fixed threshold.

A score is then computed for each extracted passage. This score takes into account three factors:

- the number and the significance of the words of the question that are present in the passage. The significance of a question word is evaluated by its normalized information, computed from 2 years of the *Le Monde* newspaper;
- the presence in the passage of a named entity that corresponds to the expected answer type when the answer type is a named entity;
- the density of the words of the question in the passage.

More precisely, the score of a passage p_i is:

$$score(p_i) = \alpha \cdot wordScore(p_i) + \beta \cdot neScore(p_i) + \gamma \cdot densityScore(p_i) \quad (1)$$

where α,β and γ are modulators[5] and all the scores are between 0.0 and 1.0.
The word score is given by:

$$wordScore(p_i) = \frac{\sum_k significance(w_k)}{number\ of\ question\ plain\ words} \quad (2)$$

where w_k is a word of p_i that is a word of the question.

The named entity score is equal to 1.0 if a named entity that corresponds to the expected answer type is present in p_i and to 0.0 otherwise. The density score is defined with respect to a reference size for a passage, given by:

$$reference\ size = 2 * activSpread + number\ of\ question\ words(p_i) \quad (3)$$

[4] Corpus Processing Tools, available at: `http://copt.sourceforge.net`

[5] For our CLEF-QA run, α and β were equal to 1.0. The value of γ depends on two factors. The first one is the core modulator value set as a parameter (equal to 1.0 in our case). This factor is modulated by :

$$\frac{number\ of\ question\ words(p_i)^4}{(number\ of\ question\ words(p_i) + 1)^4}$$

which makes the density score less important when the number of question words that are present in p_i is high.

If the size of p_i is less than *reference size*, its density score is equal to its maximal value, *i.e.* 1.0. Otherwise, it is attenuated with respect to how much the size of p_i is greater than $reference\,size$, by being equal to:

$$densityScore(p_i) = \frac{1}{\sqrt{\frac{passage\ size}{reference\,size}}} \quad (4)$$

Once their score is computed, the passages are sorted according to the decreasing order of their score et the first N passages are kept for the further steps[6].

4 From passages to answer

Oedipe was first developed as a question answering system dedicated to find passage answers rather than to find short answers. We adapted it for the CLEF-QA evaluation but without changing its overall design. Hence, it first searches for passage answers and then tries to find short answers in them.

4.1 From passages to passage answers

Oedipe locates a passage answer in each selected passage. This process consists in moving a window over the target passage and to compute a score at each position of the window according to its content. The size of this window is equal to the size of the answer to extract[7]. The extracted answer is the content of the window for its position with the higher score.

The way the window is moved depends on the expected answer type. If the expected answer is not a named entity, the window is moved over each plain word of the passage. Otherwise, it is moved only over the positions where a named entity that corresponds to the expected answer type is present. In both cases, the score computed at each position is the sum of two sub-scores:

- a score evaluating the number and the significance of the question words that are in the window. This score is the same as the word score for passages (see 2);
- a score that is directly equal to the proportion of the named entities of the question that are in the window.

For questions whose expected answer is not a named entity, it is frequent to have several adjacent positions with the same score. When such a case happens with the highest score of the passage, the selected passage answer is taken from the middle of this zone and not from its beginning, as the answer often comes after the words of the question.

Finally, as for passages, all the passage answers are sorted according to the decreasing order of their score. If the score of the highest answer is too low, *i.e.* below a fixed threshold, Oedipe assumes that there is no answer to the considered question.

4.2 From passage answers to short answers

When the expected answer is a named entity, the extraction of short answers is straightforward: the passage answer with the highest score is selected and the named entity on which the passage answer extraction window was centered is returned as a short answer. In the other case, a search for a short answer based on a small set of heuristics is performed. This search assumes that a short answer is a noun phrase. Hence, Oedipe locates all the nouns phrases of the passage answer by applying the following morpho-syntactic pattern:

[6] N is equal to 20 for this evaluation.

[7] The window size was equal to 250 characters for the CLEF-QA evaluation.

(DET|NP|NC)(NC|NP|ADJ|PREP)(ADJ|NC|NP)[8]

Then, it computes a score for each of them. This score takes into account both the size of the answer and its context:

- its base is proportional to the size of the answer, with a fixed limit;
- it is increased by a fixed value each time a specific element is found in its close context (2 words). This element can be one of the named entities of the question or more generally, an element that is characteristic of the presence of a definition, such as a period, a parenthesis or the verb "to be".

The final score of a short answer is the sum of its passage answer score and of its short answer score. The short answer with the highest score is returned as the answer to the considered question.

5 Results

We submitted only one run of the Oedipe system for the CLEF-QA 2005 evaluation. On the 200 test questions, Oedipe returned a right answer for only 28 questions (6 of them were temporally restricted factoid), 3 inexact answers and an unsupported answer for 4 other questions. Moreover, the detection of a lack of answer by Oedipe was right for only one question among the three it marked, whereas this lack had to be found for 20 questions. Table 1 gives the best results of the seven participants to the monolingual track for French. As we can see from this table, the results of Oedipe (system 7) are not good but they are not too far from the results of half of the participants.

Table 1: CLEF-QA 2005 results for the French monolingual track

systems	# right answers	# inexact answers	# unsupported answers	score with question difficulty
1	128	8	2	67.5
2	70	10	0	30.75
3	46	7	4	17.75
4	35	8	1	15.25
5	33	20	3	17.75
6	29	21	2	15
7	**28**	**3**	**4**	**16.75**

To take into account the fact that all the questions do not have the same level of difficulty, we have computed a specific score (see the last column of Table 1). The difficulty of a question is evaluated by the number of systems that do not return a right answer for it. We computed the mean (denoted M_{diff}) and the standard deviation (denoted SD_{diff}) of the difficulty values for the 200 questions[9] and set the score of a right answer to a question as follows:

$$\begin{array}{lll} score = 0.25 & if & difficulty \leq M_{diff} - SD_{diff} \\ score = 0.5 & if & difficulty \leq M_{diff} \\ score = 0.75 & if & difficulty \leq M_{diff} + SD_{diff} \\ score = 1 & if & difficulty > M_{diff} + SD_{diff} \end{array}$$

This score confirms the fact that the results of Oedipe are not very far from the results of most of the participants and that they could be improved quite quickly as it misses some "easy" questions.

[8] DET: article, NP: proper noun, NC: common noun, ADJ: adjective, PREP: preposition

[9] These difficulty values were computed from all the runs with French as a target language.

6 Conclusion

We have presented in this article the version of the Oedipe system that participated to the French monolingual track of the CLEF-QA 2005 evaluation. Its results are not very high but they are coherent with the degree of simplicity of the system. A quick analysis of its results shows that such a simple system can be sufficient to answer to around 20% of factoid questions but is totally inefficient for answering to more complex questions such as definition questions. Hence, we will focus our future work on that aspect. A short-term improvement about it could be to make use of the capabilities of LIMA about syntactic analysis for delimiting short answers instead of using an approximate pattern. In a more long-term plan, we would like to elaborate an instance-based approach for extracting short answers, which could avoid building of a set of manual patterns as it is often done.

References

[1] Christelle Ayache. Campagne EVALDA/EQUER : Evaluation en question-réponse, rapport final de la campagne EVALDA/EQUER. Technical report, ELDA, 2005.

[2] Antonio Balvet, Mehdi Embarek, and Olivier Ferret. Minimalisme et question-réponse : le système Oedipe. In *Actes de l'Atelier EQueR de la conférence TALN 2005*, pages 77–80, Dourdan, France, June 2005.

[3] Romaric Besançon and Gaël Chalendar (de). L'analyseur syntaxique de LIMA dans la campagne d'évaluation EASY. In *Actes de la 12e conférence annuelle sur le Traitement Automatique des Langues Naturelles, TALN 2005*, Dourdan, France, June 2005.

[4] Romaric Besançon, Gaël Chalendar (de), Olivier Ferret, Christian Fluhr, Olivier Mesnard, and Hubert Naets. *4^{th} Workshop of the Cross-Language Evaluation Forum, CLEF 2003*, chapter Concept-based Searching and Merging for Multilingual Information Retrieval: First Experiments at CLEF 2003, pages 174–184. 2004.

[5] Romaric Besançon, Olivier Ferret, and Christian Fluhr. LIC2M experiments at CLEF 2004. In *5^{th} Workshop of the Cross-Language Evaluation Forum, CLEF 2004*, 2004.

[6] Olivier Ferret, Brigitte Grau, Martine Hurault-Plantet, Gabriel Illouz, and Christian Jacquemin. Document selection refinement based on linguistic features for QALC, a question answering system. In *Recent Advances in Natural Language Processing (RANLP 2001)*, 2001.

[7] Dominique Laurent and Patrick Séguéla. QRISTAL, système de questions-réponses. In *TALN 2005*, pages 53–62, 2005.

[8] D. Moldovan, S. Harabagiu, R. Girju, P. Morarescu, A. Novischi, A. Badalescu, and O. Bolohan. Lcc tools for question answering. In *TREC 2002*, number 2003.

[9] Memno van Zaanen. *Boostrapping Structure into Language: Alignement-Based Learning*. PhD thesis, University of Leeds, 2001.